# Attending to General and Content-Specific Dimensions of Teaching: 

# Exploring Factors Across Two Observation Instruments 

David Blazar<br>david_blazar@mail.harvard.edu<br>David Braslow<br>Harvard Graduate School of Education<br>Charalambos Y. Charalambous<br>University of Cyprus<br>Heather C. Hill<br>Harvard Graduate School of Education

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. Additional funding comes from the National Science Foundation Grant 0918383. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors would like to thank the teachers who participated in this study.

Draft Document. Please do not circulate or cite without permission.


#### Abstract

New observation instruments used in research and evaluation settings assess teachers along multiple domains of teaching practice, both general and content-specific. However, this work infrequently explores the relationship between these domains. In this study, we use exploratory and confirmatory factor analyses of two observation instruments - the Classroom Assessment Scoring System (CLASS) and the Mathematical Quality of Instruction (MQI) - to explore the extent to which we might integrate both general and content-specific views of teaching. Importantly, bi-factor analyses that account for instrument-specific variation enable more robust conclusions than in existing literature. Findings indicate that there is some overlap between instruments, but that the best factor structures include both general and content-specific practices. This suggests new approaches to measuring mathematics instruction for the purposes of evaluation and professional development.


Draft Document. Please do not circulate or cite without permission.

## Introduction and Background

Many who study teaching and learning view it as a complex craft made up of multiple dimensions and competencies (e.g., Cohen, 2011; Lampert, 2001; Leinhardt, 1993). In particular, older (Brophy, 1986) and more recent (Grossman \& McDonald, 2008; Hamre et al., 2013) work calls on researchers, practitioners, and policymakers to consider both general and more contentspecific elements of instruction. General classroom pedagogy often includes soliciting student thinking through effective questioning, giving timely and relevant feedback to students, and maintaining a positive classroom climate. Content-specific elements include ensuring the accuracy of the content taught, providing opportunities for students to think and reason about the content, and using evidence-based best practices (e.g., linking between representations or use of multiple solution strategies in mathematics).

However, research studies and policy initiatives rarely integrate these views of teaching in practice. For example, new teacher evaluation systems often ask school leaders to utilize general instruments such as the Framework for Teaching when observing instruction (Center on Great Teachers and Leaders, 2013). Professional developments efforts focus on content-specific practices (e.g., Marilyn Burns's Math Solutions) or general classroom pedagogy (e.g., Doug Lemov's Teach Like a Champion) but infrequently attend to both aspects of teaching simultaneously. This trend also is evident in research settings, with most studies of teaching quality drawing on just one observation instrument - either general or content-specific (see, for example, Hill et al, 2008; Hafen et al., 2014; Kane, Taylor, Tyler, \& Wooten, 2011; Grossman, Loeb, Cohen, \& Wyckoff, 2013; McCaffrey, Yuan, Savitsky, Lockwood, \& Edelen, 2014; Pianta, Belsky, Vandergrift, Houts, \& Morrison, 2008).

Draft Document. Please do not circulate or cite without permission.

To our knowledge, only two analyses utilize rigorous methods to examine both general and content-specific teaching practices concurrently. Both draw on data from the Measures of Effective Teaching (MET) project, which includes scores on multiple observation instruments from teachers across six urban school districts. Using a principal components analysis framework, Kane and Staiger (2012) found that items tended to cluster within instrument to form up to three principal components: one that captured all competencies from a given instrument simultaneously, analogous to a single dimension for "good" teaching; a second that focused on classroom or time management; and a third that captured a specific competency highlighted by the individual instrument (e.g., teachers' ability to have students describe their thinking for the Framework for Teaching, and classroom climate for the Classroom Assessment Scoring System). Using the same data, McClellan and colleagues (2013) examined overlap between general and content-specific observation instruments. Factor analyses indicated that instruments did not have the same common structure. In addition, factor structures of individual instruments were not sensitive to the presence of additional instruments, further suggesting independent constructs. Without much overlap between instruments, the authors identified as many as twelve unique factors. Together, this work suggests that instruments that attend either to general or contentspecific aspects of instruction cannot sufficiently capture the multi-dimensional nature of teaching.

At the same time, these findings point to a challenge associated with looking for factors across instruments: the existence of instrument-specific variation. Due to differences in the design and implementation of each instrument - such as the number of score points or the pool of raters - scores will tend to cluster more strongly within instruments than across them (Crocker \& Algina, 2008). Therefore, distinctions made between teaching constructs - including general

Draft Document. Please do not circulate or cite without permission.
versus content-specific ones - may be artificial. However, it may be possible to account for some instrument-specific variation using bi-factor models, in which teachers' scores are explained by both instructional and method, or instrument-specific, factors (Chen, Hayes, Carver, Laurenceau, \& Zhang, 2012; Gustafsson, \& Balke, 1993).

To extend this line of work, we analyze data from a sample of fourth- and fifth-grade teachers with scores on two observation instruments: the Classroom Assessment Scoring System (CLASS), a general instrument, and the Mathematical Quality of Instruction (MQI), a contentspecific instrument. Drawing on exploratory and confirmatory factor analyses, we examine the relationship between instructional quality scores captured by these two instruments. In addition, we examine what integration of general and content-specific views of teaching might look like that is, whether teaching is the sum of all dimensions across these two instruments or whether there is a more parsimonious structure. It is important to note that, while we focus specifically on mathematics, future research may attempt to explore this issue for other content areas.

Results from this analysis can inform evaluation and development policies. If findings indicate that both general and content-specific factors are necessary to describe instructional quality, then school leaders may seek to utilize multiple instruments when viewing instruction. Evaluation scores on multiple competencies and elements of teaching may be particularly important for development efforts that seek to improve teachers' practice in specific areas.

## Data and Participants

Our sample consists of 390 fourth- and fifth-grade teachers from five school districts on the east coast of the United States. Four of the districts were part of a large-scale project from the National Center for Teacher Effectiveness focused around the collection of observation scores and other teacher characteristics. Teachers from the fifth district participated in a separate

Draft Document. Please do not circulate or cite without permission.
randomized controlled trial of a mathematics professional development program that collected similar data on teachers as the first project. Both projects spanned the 2010-11 through the 201213 school years. In the first project, schools were recruited based on district referrals and size; the study required a minimum of two teachers in each school at each of the sampled grades. Of eligible teachers in these schools, roughly $55 \%$ agreed to participate. In the second study, we only include the treatment teachers for the first two years, as observation data were not collected for the control group teachers. We have video data on teachers in both groups in the third year.

Teachers' mathematics lessons $(\mathrm{N}=2,276)$ were captured over a three-year period, with a yearly average of three lessons per teacher for the first project and six lessons per teacher for the second project. Videos were recorded using a three-camera, unmanned unit; site coordinators turned the camera on prior to the lesson and off at its conclusion. Most lessons lasted between 45 and 60 minutes. Teachers were allowed to choose the dates for capture in advance and were directed to select typical lessons and exclude days on which students were taking a test. Although it is possible that these videotaped lessons are different from teachers' general instruction, teachers did not have any incentive to select lessons strategically as no rewards or sanctions were involved with data collection. In addition, analyses from the MET project indicate that teachers are ranked almost identically when they choose lessons to be observed compared to when lessons are chosen for them (Ho \& Kane, 2013).

Trained raters scored these lessons on two established observation instruments: the CLASS, which focuses on general teaching practices, and the MQI, which focuses on mathematics-specific practices. Validity studies have shown that both instruments successfully capture the quality of teachers' instruction, and specific dimensions from each instrument have been shown to relate to student outcomes (Blazar, 2015; Hill, Charalambous, \& Kraft, 2012; Bell,

Draft Document. Please do not circulate or cite without permission.

Gitomer, McCaffrey, Hamre, \& Pianta, 2012; Kane \& Staiger, 2012; Pianta et al., 2008). For the CLASS, one rater watched each lesson and scored teachers' instruction on 12 items for each fifteen-minute segment on a scale from Low (1) to High (7). For the MQI, two raters watched each lesson and scored teachers' instruction on 13 items for each seven-and-a-half-minute segment on a scale from Low (1) to High (3) (see Table 1 for a full list of items and descriptions).

We exclude from this analysis a single item from the MQI, Classroom Work is Connected to Math, as it is scored on a different scale (Not True [0] True [1]) and did not load cleanly onto any of the resulting factors. One item from the CLASS (Negative Climate) and three from the MQI (Major Errors, Language Imprecisions, and Lack of Clarity) have a negative valence. For both instruments, raters had to complete an online training, pass a certification exam, and participate in ongoing calibration sessions. Separate pools of raters were recruited for each instrument.

We used these data to create three datasets. The first is a segment-level dataset that captures the original scores assigned to each teacher by raters while watching each lesson. ${ }^{1}$ The second is a lesson-level dataset with scores for each item on both the CLASS and MQI averaged across raters (for the MQI) and segments. The third is a teacher-level dataset with scores averaged across lessons. For most analyses that we describe below, we fit models using all three datasets. ${ }^{2}$ However, we focus our discussion of the results using findings from our teacher-level data for three reasons. First and foremost, our constructs of interest (i.e., teaching quality) lie at the teacher level. Second, patterns of results from these additional analyses (available upon

[^0]Draft Document. Please do not circulate or cite without permission.
request) lead us to substantively similar conclusions. Finally, other similar studies also use teachers as the level of analysis (Kane \& Staiger, 2012; McClellan, Donoghue, \& Park, 2013).

## Analysis Strategy

To explore the relationship between general and content-specific elements of teaching, we conducted three sets of analyses. We began by examining pairwise correlations of items across instruments. This allowed us to explore the degree of potential overlap in the dimensions of instruction captured by each instrument.

Next, we conducted a set of exploratory factor analyses (EFA) to identify the number of factors we might expect to see, both within and across instruments. In running these analyses, we attempted to get parsimonious models that would explain as much of the variation in the assigned teaching quality ratings with as few factors as possible. We opted for non-orthogonal rotations (i.e., direct oblimin rotation), which assumes that the extracted factors are correlated. We did so given theory (Hill, 2010; Learning Mathematics for Teaching, 2011; Pianta \& Hamre, 2009) and empirical findings (Hill et al., 2008; Pianta, Belsky, Vandergrift, Houts, \& Morrison, 2008) suggesting that the different constructs within each instrument are inter-correlated. ${ }^{3}$

While we conducted these EFA to look for cross-instrument factors, prior research suggests that we would not expect to see much overlap across instruments (McClellan et al., 2013). Therefore, we used confirmatory factor analysis (CFA) to account for construct-irrelevant variation caused by the use of the two different instruments. In particular, we utilized bi-factor

[^1]Draft Document. Please do not circulate or cite without permission.
models (Chen et al., 2012) to extract instrument-specific variation, and then tested factor structures that allowed items to cluster across instruments.

Our use of CFA is non-traditional. Generally, CFA attempts to find models that achieve adequate global fit by building successively more complex models. As we are interested in parsimonious models that might not fully capture the observed data, and because there are a number of features of our data that are not included in our model (e.g., use of multiple raters), we instead look at incremental improvements in fit indices to evaluate different teacher-level instructional factor structures.

## Results

Our correlation matrix shows that some items on the CLASS and MQI are moderately related at the teacher level (see Table 2). For example, both Analysis and Problem Solving and Instructional Dialogue from CLASS are correlated with multiple items from the MQI (Mathematical Language, Use of Student Productions, Student Explanations, Student Mathematical Questioning and Reasoning, and Enacted Task Cognitive Activation) above 0.30. Three items from the MQI - Mathematical Language, Use Student Productions, and Student Mathematical Questioning and Reasoning (SMQR) - are correlated with multiple items from CLASS at similar magnitudes. The largest observed cross-instrument correlation of 0.41 is between Analysis and Problem Solving and Use Student Productions. Even though we run 156 separate tests, the 104 statistically significant correlations are much higher than the $5 \%$ we would expect to see by chance alone. These findings suggest that items from the two instruments seem to be capturing somewhat similar facets of instruction. Therefore, factor structures might include factors with loadings across instruments.

Draft Document. Please do not circulate or cite without permission.

At the same time, there do appear to be distinct elements of instruction captured by each instrument. In particular, the three items capturing mathematical errors - embedded deeply in a content-specific view of teaching - are not related to items from CLASS, suggesting that this might be a unique construct from more general elements of instruction or classroom pedagogy. Further, five items from the CLASS correlate with items from the MQI no higher than 0.3.

Next, we present results from the EFA. First we note that the Kaiser-Mayer-Olkin (KMO) value in all factor analyses exceeded the acceptable threshold of meritorious values ( 0.80 ), thus suggesting that the data lent themselves to forming groups of variables, namely, factors (Kaiser, 1974). Initial results point to six factors with eigenvalues above 1.0, a conventionally used threshold for selecting factors (Kline, 1994); scree plot analysis also support these six as unique factors (Hayton, Allen, \& Scarpello, 2004). However, even after rotation, no item loads onto the sixth factor at or above 0.4 , which is often taken as the minimum acceptable factor loading (Field, 2013; Kline, 1994). Two considerations guide our decision regarding which of the more parsimonious models best fit our data: the percent of variance explained by each additional factor, and the extent to which the factors have loadings that support substantive interpretations. We discard the five-factor solution given that it only explains three percent more of the variance in our data, and therefore contributes only minimally to explaining the variance in the assigned teaching quality ratings (Field, 2013; Tabachnick \& Fidell, 2001). In addition, items that load onto the fifth factor almost all cross load onto other factors; generally, these loadings are weak. We also exclude one- and two-factor solutions, as neither explains more than $50 \%$ of variation in our data, a conventionally agreed-upon threshold for accepting a factor structure (Kline, 1994).

Draft Document. Please do not circulate or cite without permission.

In Tables 3a and 3b, we present eigenvalues, percent of variance explained, and factor loadings for a parsimonious list of factors generated from the remaining three- and four-factor solutions. In the three-factor solution, 22 of the 25 items load clearly onto only one factor. Of the remaining three items, Analysis and Problem Solving from the CLASS loads strongly onto the first factor and also has a notable loading on the second factor. Two items from the MQI, Mathematical Language and Mathematical Generalizations, have communalities that are considerably low. In addition, Mathematical Language has loadings on both the first and second factors of similar magnitudes, both below the acceptable threshold of 0.4 . Together, these three factors explain roughly $53 \%$ of the variance in our data, and all have acceptable reliability indices. Based on the patterns of item loadings, we label these three factors "Ambitious General Instruction" (with all 12 items from the CLASS instrument, Cronhach's alpha $=0.91$ ), "Ambitious Mathematics Instruction" (with 10 items from the MQI, Cronhach's alpha $=0.87$ ), and "Mathematical Errors" (with the last three items from the MQI, Cronhach's alpha $=0.76$ ). The first two factors are correlated at 0.33 , and the latter two factors are correlated at -0.23 ; the correlation between "Ambitious General Instruction" and "Mathematical Errors" is negligible ( $r$ $=0.03$ ) (see Table 4a). These correlations are in the expected directions, supporting our substantive interpretations.

When we add a fourth factor, items from CLASS split into two dimensions. One of these is substantively similar to the factor described above, which we continue to refer to as "Ambitious General Instruction" (Cronbach's alpha=0.94). The other consists of three items Behavior Management, Productivity, and Negative Climate - that instrument developers refer to as "Classroom Organization" (Cronbach's alpha $=0.82$ ). MQI items load substantively onto the same two factors described above. While we explain five percent more variation compared to a

Draft Document. Please do not circulate or cite without permission.
three-factor solution, for a total of $58 \%$, the low communality values for Mathematical Language and Generalizations persist. Additionally, a number of items from the CLASS cross load onto both "Ambitious General Instruction" and "Classroom Organization." Some of these items have additional loadings on "Ambitious Mathematics Instruction" above 0.3. The two CLASS factors are correlated most strongly, at 0.51 (see Table 4b). Similar to above, "Ambitious General Instruction" and "Ambitious Mathematics Instruction" are correlated at $r=0.35$, and this latter factor is correlated with "Mathematical Errors" at $r=-0.24$. Other correlations are below $r=$ 0.20. Although the four-factor solution explains a somewhat higher percentage of the total variance in the data than the three-factor solution, this solution has more cross loadings than the simpler solution.

The cross loadings from the EFA suggest that some shared variation exists across instruments, even though the suggested factors are largely within-instrument. To explore this further, we proceed to CFA to test whether extracting instrument-specific variation leads us to one of the two solutions described above, or to another solution. We present the structure of these theory-driven models in Tables 5a and 5b, which document non-bi-factor and bi-factor models, respectively. Models 1 through 4 are non-bi-factor models. Models 3 and 4 correspond to the three- and four-factor solutions from the EFA analyses, with items restricted to load only on their primary factors from the EFA. These models provide a basis for comparison with the more complex models. We also run models with just one instructional factor (Model 1) and two factors comprised of items from each instrument (Model 2) in order to examine whether the suggested models that emerged from the EFA have better fit as compared to that of more parsimonious models. This is a common practice when running CFA (Kline, 2011). Models 5 through 8 are bi-factor models, each with two method factors ("CLASS" and "MQI") that

Draft Document. Please do not circulate or cite without permission.
attempt to extract instrument-specific variation, as well as varying numbers of instructional factors including cross-instrument factors. In these models, items are specified to load on one substantive factor and on one instrument factor. Model 5 includes just one substantive factor, "Ambitious Instruction," for a total of three factors (one substantive and two instrument factors). Models 6 and 7 each has two broad substantive factors, "Ambitious Instruction" and one additional specific factor from either the CLASS or MQI instrument that the EFA suggested were important (i.e., "Classroom Organization" or "Mathematical Errors"). Finally, Model 8 includes all three of these substantive factors. Notably, we do not include any bi-factor models with more than three substantive factors. This is because we hypothesize that bi-factor models might lead to a more parsimonious solution than non-bi-factor models by allowing for crossinstrument factors.

In Table 6, we present fit indices for each of these models using robust maximum likelihood estimation to account for the non-normality of some items. We present standard fit statistics and identify the best-fitting models by comparing AICs and BICs (Sivo, Fan, Witta, \& Willse, 2006). For models that are nested, we also can test formally for differences in model fit.

Of the non bi-factor models, Model 4 appears to have the best fit. We test formally for difference in fit between Model 4 and Model 3, which has the next lowest AIC and BIC statistics. Due to the use of robust maximum likelihood estimation, we use Satorra-Bentler Scaled Chi-Squared to compare the fit of nested models (Satorra \& Bentler, 1999). Despite three additional parameters in Model 4 compared to Model 3, the former model significantly reduces the overall adjusted model chi-square $\left(\Delta \chi_{d f=3, N=390}^{2}=144.67, p<0.001\right)$, thus improving model fit. Therefore, we conclude that, of the non bi-factor models, Model 4 is the best fit to the data.

Draft Document. Please do not circulate or cite without permission.

Similarly, we compare fit indices for bi-factor models and find that Model 8 has the best fit. AIC and BIC values are substantially lower than the next best model, Model 6. Formally comparing Model 8 to Model 6, we find evidence that the former is a better fit to the data than the latter $\left(\Delta \chi_{d f=2, N=390}^{2}=39.82, p<0.001\right)$.

Of the remaining two models - Models 4 and 8 - we cannot test formally for differences in model fit, given that they are not nested. In addition, we argue that direct comparisons may not be appropriate, given that Model 8 has an additional factor. Rather, examinations of the factor loadings of these two models provide useful insight (see Tables 7a and 7b). As results from the EFA, we find substantive support for Model 4, where items have statistically significant loadings on their respective factors, generally above 0.4 . Two item loadings for Generalizations and Mathematical Language fall just below this threshold; however, we observed similar issues in the EFA.

Factor loadings in Model 8 are less clean. We hypothesized that variation in a particular item should be accounted for both by an instrument factor and to the content of the item. This is true for the "Classroom Organization" factor, where all three items have loadings on both the instrument and the substantive factor above 0.4. Four items from the MQI specified to load on the "Ambitious Instruction" factor also meet this condition. However, for "Mathematical Errors", all three items load only onto the substantive factor at this threshold; loadings on the instrument factor are below 0.4 , even though they are statistically significant. Conversely, for many items included in the "Ambitious Instruction" factor, almost all of the variation loads onto the instrument factor. For example, Linking and Connections has a strong loading of 0.57 on the "MQI" instrument factor but a non-significant loading of 0.14 on the substantive "Ambitious Instruction" factor. Items from the CLASS specified to load on this same factor have loadings no

Draft Document. Please do not circulate or cite without permission.
higher than 0.35 . One reason for this may be that there is a large degree of overlap of items between this substantive factor and the two instrument factors; that is, we specify that all but three CLASS items and all but three MQI items load onto the "Ambitious Instruction" factor. Another explanation is that our sample size is small relative to recommended guidelines for stable parameter estimates. The rule of thumb is to have five to 10 observations per parameter estimated (Kline, 2011), yet we only have 390 observations for approximately 100 parameters.

That said, two pieces of evidence suggest continued consideration of Model 8 as a plausible factor structure. First, comparison of Model 5 to Models 1 and 2 indicates that including both substantive and instrument-specific factors helps describe the variation in our data. Above, we noted that we could not test formally for differences in model fit between Models 4 and 8, given that they are not nested. However, Models 1 and 2 - non-bi-factor models - are nested within Model 5 - a bi-factor model. Model 5 includes one substantive factor "Ambitious Instruction" - and two instrument-specific factors - "CLASS" and "MQI". Model 1 includes the former factor alone, while Model 2 only includes the latter two factors. In both cases, we find that Model 5 is a better fit to the data (compared to Model 1: $\Delta \chi_{d f=26, N=390}^{2}=$ 1236.09, $p<0.001$; compared to Model 2: $\Delta \chi_{d f=25, N=390}^{2}=100.64, p<0.001$ ). This suggests that a model that includes both instrument and substantive factors is a better fit than a model that only includes one or the other. Second, even though factor loadings on Model 8 do not load onto their specified factors at conventional levels, many are still statistically significant. In light of the limitations of our bi-factor models that we describe above, we may consider a lower threshold than 0.4 for factor loadings. In this case, patterns in Model 8 are more consistent with theory, with items loading both onto a substantive and instrument-specific factor. In turn, this implies

Draft Document. Please do not circulate or cite without permission.
that variation in these items cannot be captured by considering an instrument factor alone and that a substantive factor also needs to be considered.

Comparing Model 8 to Model 4 leads us to note three trends. First, it appears that both general and content-specific dimensions are needed to describe variation across teachers. In both of these models, "Mathematical Errors" and "Classroom Organization" form their own factors, even though this was not true in the three-factor solution from the EFA. In that analysis, we found that items from "Classroom Organization" clustered with the "Ambitious General Instruction" factor.

Second, we have some suggestive evidence for overlap between instructional components of the CLASS and MQI instruments. This is most evident for items related to students' cognitive engagement in class and teachers' interactions with students around the content. In Model 8, four items from the MQI - Use Productions, Student Explanations, Student Mathematical Questioning and Reasoning, and Enacted Task Cognitive Activation - load onto the "Ambitious Instruction" factor that we specify to include items from both instruments. If we consider a slightly lower threshold for factor loadings around 0.3 , then three related items from the CLASS instrument - Respect for Student Perspectives, Analysis and Problem Solving, and Instructional Dialogue - also appear to load onto this same substantive factor.

At the same time, findings from Model 8 are less clear about other items from the CLASS and MQI specified to load onto this common factor. In particular, two items from CLASS - Positive Climate and Content Understanding - and five items from the MQI - Linking and Connections, Explanations, Generalizations, Language, and Remediation - have nonsignificant loadings below 0.2 , which suggests that most of their variance appears to be captured by the instrument-factor. Interestingly, five of these six items (excluding Positive Climate) are

Draft Document. Please do not circulate or cite without permission.
rooted in a content-specific view of instruction. Therefore, it is possible that these items might form a separate cross-instrument factor analogous to our "Ambitious Mathematics Instruction" in Model 4. We did not explore this as a possible factor structure, as this would lead us to the same general conclusion as Model 4, with "Ambitious Instruction" splitting into two substantive factors, "Ambitious General Instruction" and "Ambitious Mathematics Instruction."

Finally, we note that none of our models meet traditional fit criteria. Therefore, in future studies, additional factors - possibly drawing on additional observation instruments - would be needed to fully explain the variation across teachers. However, the three or four factors identified in Models 4 and 8 seem to account for a sizable amount of that variation.

## Discussion and Conclusion

Results from this study indicate that, indeed, integrating both general and content-specific views of teaching provide a more complete picture than focusing on just one. Although we find some overlap between elements of instruction captured by the CLASS and MQI instruments captured in the "Ambitious Instruction" factor in Model 8 - we also find strong evidence for factors that are distinct to each; "Mathematical Errors" is content specific and "Classroom Organization" is more general. This is true even when we extract instrument-specific variation. As such, these findings provide empirical support for the argument that, when studying complex phenomena such as that of teaching, we need to take both types of factors into consideration if we are to better understand and capture the quality of instruction experienced by students in the classroom.

At the same time, while our general conclusions align with related work from the MET study, we also note important areas of disagreement. Specifically, our analyses demonstrate support for between three and four instructional factors to describe variation across teachers, far

Draft Document. Please do not circulate or cite without permission.
fewer than the number identified by McClellan and colleagues (2013) who also examine factor structures using multiple observation instruments. One reason for this likely is the fact that the MET data include scores from five observation instruments, while ours include scores from two. It is possible that we might find more factors if we were to score the same instruction on additional instruments. Another plausible reason explaining the discrepancy between the MET findings and ours might be due to the models tested in each study: unlike in the MET study, our work used bi-factor models that have the potential to account for any instrument-related variance, thus reducing the number of unique factors. Future research is needed to identify the underlying dimensionality of instruction and how many total factors are needed to capture variability in practices across teachers.

We believe that our findings have important implications for theory, policy, and practice. First, with regard to theory, our findings align with older (Brophy, 1986) and more recent (Grossman \& McDonald, 2008; Hamre et al., 2013) calls to integrate general and more contentspecific perspectives when describing instructional quality. As such, our findings imply a need to develop theoretical frameworks that are more comprehensive and encompass both types of perspectives. Doing so will require a much closer collaboration among scholars from projects representing different teaching perspectives.

Second, the multidimensional nature of instruction - including both general and contentspecific practices - requires evaluation systems that reflect this complex structure. Current processes that assess teachers on just one instrument likely mask important variability within teachers. For example, a teacher who scores very low on one dimension of the instrument used to conduct the evaluation might score much higher on a dimension not included in the instrument. Job decisions may be made without a full picture of that teachers' effectiveness. Issues also arise

Draft Document. Please do not circulate or cite without permission.
from evaluation systems that come to one overall rating by averaging across dimensions within a given instrument, as often occurs in practice (Center on Great Teachers and Leaders, 2013). For example, a teacher who scores high on classroom climate but low on classroom organization may earn an overall score in the middle of the distribution even though he or she is someone who might be an appropriate target for professional development or possibly for removal. Future research may explore the optimal weight that each factor might have when attempting to develop instruments that capture both general and content-specific elements of instruction. In addition, work may explore ways to combine these two perspectives in observational systems that are designed for cost effectiveness.

Third, by attending to both general and content-specific dimensions of teaching in evaluation processes, education agencies may be better able to target support to teachers or recognize teachers based on their weak or strong skill sets. Specifically, teachers could receive distinct dimension-specific scores that lead to individualized support targeted at skills and areas where they are lacking. Without this sort of information, evaluation scores may only be useful for generalized one-size-fits all professional development, which has not proven effective at increasing teachers' instructional quality or student achievement (Hill, 2007; Yoon, Duncan, Lee, Scarloss, \& Shapley, 2007), or for dismissal or promotion.

Finally, the results of this study also have practical implications for evaluative raters in the process of scoring teachers' instruction. Even though prior work highlights the ability of principals, peers, and other school leaders to accurately identify teachers who are effective at raising student achievement (Jacob and Lefgren, 2008; Rockoff \& Speroni, 2010; Rockoff, Staiger, Kane, \& Taylor, 2012), other work indicates that specific types of instruction particular in a content area - require raters attune to these elements. For example, Hill and
colleagues (2012) show that raters who are selectively recruited due to a background in mathematics or mathematics education and who complete initial training and ongoing calibration score more accurately on the MQI than those who are not selectively recruited. Therefore, calls to identify successful teachers through evaluations that are "better, faster, and cheaper" (Gargani \& Strong, 2014) may not prove useful across all instructional dimensions.

Current efforts to evaluate teachers using multiple measures of teacher and teaching effectiveness are an important shift in the field. Evaluations can serve as an effective resource for teachers and school leaders, as long as they take into account the underlying dimensionality of teaching practice that currently exists in classrooms. In this study, we provide evidence underscoring the importance of working at the intersection of both general and content-specific practices. Continued research is needed to understand more fully the true dimensionality of teaching and how these dimensions, in isolation or in conjunction, contribute to student learning.

Draft Document. Please do not circulate or cite without permission.

## References

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., \& Pianta, R. C. (2012). An argument approach to observation protocol validity. Educational Assessment, 17(2-3), 62-87.

Blazar, D. (2014). The effect of high-quality mathematics instruction on student achievement: Exploiting within-school, between-grade, and cross-cohort variation from observation instruments. Working Paper.

Brophy, J. (1986). Teaching and learning mathematics: Where research should be going? Journal for Research in Mathematics Education, 17 (5), 323-346.

Center on Great Teachers and Leaders (2013). Databases on state teacher and principal policies. Retrieved from: http://resource.tqsource.org/stateevaldb.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J-P., \& Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. Journal of Personality, 80(1), 219-251.

Cohen, D. K. (2011). Teaching and its predicaments. Cambridge, MA: Harvard University Press.
Crocker, L., \& Algina, J. (2008). Introduction to classical and modern test theory. Mason, OH: Cengage Learning.

Field, A. (2013). Discovering statistics using IBM SPSS statistics (4 $4^{\text {th }}$ ed.). London: SAGE publications.

Gargani, J., \& Strong, M. (2014). Can we identify a successful teacher better, faster, and cheaper? Evidence for innovating teacher observation systems. Journal of Teacher Education, 65 (5), 389-401.

Grossman, P., Loeb, S., Cohen, J., \& Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added. American Journal of Education, 199(3), 445-470.

Grossman, P., \& McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. American Educational Research Journal, 45, 184-205.

Gustafsson, J., \& Balke, G. (1993). General and specific abilities as predictors of school achievement. Multivariate Behavioral Research, 28, 407-434.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., \& Pianta, R. C. (2014). Teaching through interactions in secondary school classrooms revisiting the factor structure and practical application of the classroom assessment scoring systemsecondary. The Journal of Early Adolescence. Advance online publication.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., et al. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. The Elementary School Journal, 113(4), 461-487.

Hayton, J. C., Allen, D. G., \& Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. Organizational Research Methods, 7(2), 191-205.

Hill, H. C. (2007). Learning in the teacher workforce. Future of Children, 17(1), 111-127.
Hill, H. C. (2010, May). The Mathematical Quality of Instruction: Learning Mathematics for Teaching. Paper presented at the 2010 annual meeting of the American Educational Research Association, Denver, CO.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., \& Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. Cognition and Instruction, 26(4), 430-511.

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A., Litke, E., \& Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. Educational Assessment, 17(2-3), 88-106.

Hill, H. C., Charalambous, C. Y., \& Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. Educational Researchers, 41(2), 56-64.

Ho, A. D., \& Kane, T. J. (2013). The reliability of classroom observations by school personnel. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Jacob B. A., \& Lefgren L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. Journal of Labor Economics, 20(1), 101136.

Kaiser, H. (1974). An index of factorial simplicity. Psychometrika, 39, 31-36.
Kane, T. J., Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper. Seattle, WA: Measures of Effective Teaching Project, Bill and Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., \& Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. Journal of Human Resources, 46(3), 587-613.

Kline, P. (1994). An easy guide to factor analysis. London: Routledge.
Kline, R. B. (2011). Principles and practice of structural equation modeling. New York: The Guilford Press.

Draft Document. Please do not circulate or cite without permission.

Lampert, M. (2001). Teaching problems and the problems of teaching. Yale University Press.
Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. Journal of Mathematics Teacher Education, 14, 25-47.

Leinhardt, G. (1993). On teaching. In R. Glaser (Ed.), Advances in instructional psychology (Vol.4, pp. 1-54). Hillsdale, NJ: Lawrence Erlbaum Associates.

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., \& Edelen, M. O. (2014). Uncovering multivariate structure in classroom observations in the presence of rater errors. Educational Measurement: Issues and Practice. Advance online publication.

McClellan, C., Donoghue, J., \& Park, Y. S. (2013). Commonality and uniqueness in teaching practice observation. Clowder Consulting.

Pianta, B., Belsky, J. Vandergrift, N. Houts, R., \& Morrison, F. (2008) Classroom effects on children's achievement trajectories in elementary school. American Educational Research Journal, 45(2), 365-387.

Pianta, R., \& Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. Educational Researcher, 38 (2), 109-119.

Pianta, R. C., Hamre, B. K., \& Mintz, S. (2010). Classroom Assessment Scoring System (CLASS) Manual: Upper Elementary. Teachstone.

Rockoff, J. E., \& Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. American Economic Review, 261-266.

Rockoff, J. E., Staiger, D. O., Kane, T.J., \& Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. American Economic Review, 102(7), 3184-3213.

Satorra, A., \& Bentler, P. M. (1999). A scaled difference Chi-square test statistic for moment structure analysis. Retrieved from http://statistics.ucla.edu/preprints/uclastat-preprint1999:19.

Sivo, S. A., Fan, X., Witta, E. L., Willse , J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. The Journal of Experimental Education, 74(3), 267-288.

Tabachnick, B. G., \& Fidell, L. S. (2001). Using multivariate statistics (4th ed.). New York: Harper Collins.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., \& Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement. Washington, DC: U.S. Department of Education, Institute of Education Sciences,

Draft Document. Please do not circulate or cite without permission.

National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

# Draft Document. Please do not circulate or cite without permission. 

## Tables

Table 1

| Item Descriptions |  |
| :--- | :--- |
| Items |  |
| CLASS |  |
| Negative Climate | Negative climate reflects the overall level of negativity among teachers and students in the |
| class. |  |

[^2]Table 2
Item Corre

| Items | Linking and Connections | Explanations | Multiple Methods | Generalizations | Mathematical Language | Remediation | Use Student Productions | Student Explanations | SMQR | ETCA | Major Errors | Language Imprecisions | Lack of Clarity |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Negative Climate | -0.104* | -0.104* | -0.078 | -0.103* | -0.223*** | -0.007 | -0.146** | -0.068 | -0.090~ | -0.127* | 0.001 | -0.072 | 0.027 |
| Behavior Management | 0.103* | 0.141** | 0.022 | 0.114* | 0.299*** | 0.067 | 0.182*** | 0.107* | 0.134** | 0.146** | 0.058 | 0.102* | 0.027 |
| Productivity | 0.155** | 0.215*** | 0.059 | 0.199*** | 0.299*** | 0.169*** | 0.203*** | $0.128 *$ | 0.159** | 0.216*** | 0.056 | 0.079 | 0.049 |
| Student Engagement | 0.078 | 0.112* | -0.011 | 0.089~ | 0.208*** | 0.051 | 0.228*** | 0.148** | 0.143** | 0.174*** | -0.023 | 0.03 | -0.01 |
| Positive Climate | 0.099 | 0.132** | 0.035 | 0.104* | 0.257*** | 0.049 | 0.168*** | 0.094~ | 0.150** | 0.147** | 0.005 | 0.045 | -0.014 |
| Teacher Sensitivity | 0.170*** | 0.293*** | 0.136** | 0.163** | 0.281*** | 0.208*** | 0.314*** | 0.273*** | 0.249*** | 0.305*** | -0.049 | 0.058 | -0.013 |
| Respect for Student Perspectives | 0.163** | 0.201*** | 0.203*** | 0.129* | 0.198*** | 0.140** | 0.340*** | 0.257*** | 0.262*** | 0.313*** | 0.037 | 0.016 | 0.018 |
| Instructional Learning Formats | 0.112* | 0.190*** | 0.056 | 0.147** | 0.215*** | 0.104* | 0.287*** | 0.223*** | 0.186*** | 0.268*** | -0.036 | -0.029 | -0.024 |
| Content Understanding | 0.195*** | 0.270*** | 0.062 | 0.267*** | 0.351*** | 0.176*** | 0.232*** | 0.175*** | 0.199*** | 0.219*** | 0.069 | 0.071 | 0.038 |
| Analysis and Problem Solving | 0.308*** | 0.338*** | 0.313*** | 0.165** | 0.295*** | 0.220*** | 0.410*** | 0.342*** | 0.340*** | 0.324*** | 0.003 | 0.087~ | -0.012 |
| Quality of Feedback | 0.184*** | 0.237*** | 0.113* | 0.241*** | 0.248*** | 0.218*** | 0.303*** | 0.221*** | 0.208*** | 0.296*** | 0.052 | 0.024 | 0.015 |
| Instructional Dialogue | 0.247*** | 0.264*** | 0.208*** | 0.185*** | 0.308*** | 0.192*** | 0.393*** | 0.322*** | 0.305*** | 0.321*** | 0.002 | -0.002 | 0.000 |

Draft Document. Please do not circulate or cite without permission.

Table 3a
Exploratory Factor Analyses Loadings for a Three-Factor Solution

|  | Factor 1 | Factor 2 | Factor 3 |  |
| :--- | :---: | :---: | :---: | :---: |
| Eigenvalues | 8.493 | 4.019 | 1.939 | Communalities |
| Cumulative Percent of Variance Explained | 32.32 | 46.67 | 52.95 |  |
| CLASS |  |  |  |  |
| Negative Climate | -0.578 | -0.110 | -0.003 | 0.343 |
| Behavior Management | 0.597 | 0.141 | 0.045 | 0.360 |
| Productivity | 0.691 | 0.218 | 0.059 | 0.478 |
| Student Engagement | 0.717 | 0.166 | -0.001 | 0.522 |
| Positive Climate | 0.806 | 0.165 | 0.030 | 0.662 |
| Teacher Sensitivity | 0.852 | 0.330 | -0.016 | 0.730 |
| Respect for Student Perspectives | 0.761 | 0.343 | 0.062 | 0.592 |
| Instructional Learning Formats | 0.687 | 0.253 | -0.035 | 0.475 |
| Content Understanding | 0.832 | 0.289 | 0.082 | 0.696 |
| Analysis and Problem Solving | 0.711 | 0.459 | 0.052 | 0.570 |
| Quality of Feedback | 0.812 | 0.329 | 0.059 | 0.667 |
| Instructional Dialogue | 0.841 | 0.410 | 0.031 | 0.729 |
| MQI |  |  |  |  |
| Linking and Connections | 0.199 | 0.556 | -0.190 | 0.314 |
| Explanations | 0.261 | 0.809 | -0.236 | 0.657 |
| Multiple Methods | 0.119 | 0.549 | -0.151 | 0.307 |
| Generalizations | 0.209 | 0.394 | -0.098 | 0.162 |
| Mathematical Language | 0.352 | 0.363 | -0.138 | 0.199 |
| Remediation | 0.167 | 0.609 | -0.306 | 0.400 |
| Use of Student Productions | 0.332 | 0.889 | -0.184 | 0.792 |
| Student Explanations | 0.236 | 0.808 | -0.123 | 0.658 |
| SMQR | 0.254 | 0.701 | -0.013 | 0.515 |
| ETCA | 0.296 | 0.839 | -0.236 | 0.707 |
| Major Errors | 0.011 | -0.195 | 0.835 | 0.698 |
| Language Imprecisions | 0.058 | -0.172 | 0.509 | 0.267 |
| Lack of Clarity | -0.005 | -0.174 | 0.858 | 0.739 |
|  |  |  |  |  |

Notes: Extraction method is Principal Axis Factoring. Rotation method is Oblimin with Kaiser Normalization. Cells are highlighted to identify substantive factors and potential cross-loadings (i.e., loadings on two factors of similar magnitude).

Draft Document. Please do not circulate or cite without permission.

Table 3b
Exploratory Factor Analyses Loadings for a Four-Factor Solution

|  | Factor 1 | Factor 2 | Factor 3 | Factor 4 |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Eigenvalues | 8.493 | 4.019 | 1.939 | 1.479 | Communalities |
| Cumulative Percent of Variance Explained | 32.560 | 47.036 | 53.334 | 58.063 |  |
| CLASS |  |  |  |  |  |
| Negative Climate | -0.459 | -0.122 | -0.005 | -0.687 | 0.489 |
| Behavior Management | 0.428 | 0.163 | 0.067 | 0.930 | 0.876 |
| Productivity | 0.572 | 0.232 | 0.065 | 0.772 | 0.646 |
| Student Engagement | 0.650 | 0.167 | -0.011 | 0.606 | 0.528 |
| Positive Climate | 0.803 | 0.151 | 0.005 | 0.504 | 0.679 |
| Teacher Sensitivity | 0.815 | 0.325 | -0.034 | 0.611 | 0.719 |
| Respect for Student Perspectives | 0.850 | 0.320 | 0.031 | 0.302 | 0.747 |
| Instructional Learning Formats | 0.656 | 0.249 | -0.050 | 0.492 | 0.468 |
| Content Understanding | 0.819 | 0.279 | 0.060 | 0.544 | 0.693 |
| Analysis and Problem Solving | 0.784 | 0.443 | 0.025 | 0.292 | 0.664 |
| Quality of Feedback | 0.851 | 0.311 | 0.030 | 0.426 | 0.725 |
| Instructional Dialogue | 0.896 | 0.392 | 0.000 | 0.416 | 0.811 |
| MQI |  |  |  |  |  |
| Linking and Connections | 0.212 | 0.557 | -0.194 | 0.101 | 0.314 |
| Explanations | 0.267 | 0.816 | -0.238 | 0.158 | 0.671 |
| Multiple Methods | 0.162 | 0.546 | -0.157 | -0.021 | 0.309 |
| Generalizations | 0.198 | 0.398 | -0.099 | 0.160 | 0.169 |
| Mathematical Language | 0.309 | 0.370 | -0.140 | 0.325 | 0.221 |
| Remediation | 0.181 | 0.611 | -0.308 | 0.075 | 0.401 |
| Use of Student Productions | 0.359 | 0.889 | -0.191 | 0.155 | 0.792 |
| Student Explanations | 0.273 | 0.806 | -0.129 | 0.070 | 0.656 |
| SMQR | 0.277 | 0.701 | -0.018 | 0.114 | 0.516 |
| ETCA | 0.316 | 0.841 | -0.241 | 0.148 | 0.710 |
| Major Errors | 0.018 | -0.199 | 0.835 | 0.005 | 0.697 |
| Language Imprecisions | 0.042 | -0.171 | 0.513 | 0.084 | 0.273 |
| Lack of Clarity | 0.006 | -0.177 | 0.860 | -0.013 | 0.742 |
| Nas: Exrynnnn |  |  |  |  |  |

Notes: Extraction method is Principal Axis Factoring. Rotation method is Oblimin with Kaiser Normalization. Cells are highlighted to identify substantive factors and potential cross-loadings (i.e., loadings on two factors of similar magnitude).

Table 4a
Correlations Among the Three Factors Emerging from the Exploratory Factor Analysis

| Factor | Ambitious <br> General <br> Instruction | Ambitious <br> Mathematics <br> Instruction | Mathematical <br> Errors |
| :--- | :---: | :---: | :---: |
| Ambitious General Instruction | 1.00 |  |  |
| Ambitious Mathematics Instruction | 0.33 | 1.00 |  |
| Mathematical Errors | 0.03 | -0.23 | 1.00 |

Table 4b
Correlations Among the Four Factors Emerging from the Exploratory Factor Analysis

| Factor | Ambitious <br> General <br> Instruction | Ambitious <br> Mathematics <br> Instruction | Classroom <br> Organization | Mathematical <br> Errors |
| :--- | :---: | :---: | :---: | :---: |
| Ambitious General Instruction | 1.00 |  |  |  |
| Ambitious Mathematics Instruction | 0.35 | 1.00 |  |  |
| Classroom Organization | 0.51 | 0.15 | 1.00 |  |
| Mathematical Errors | 0.02 | -0.24 | 0.01 | 1.00 |

Table 5a
Confirmatory Factor Analysis Model Organization for Non-Bi-Factor Models

| Items | Model 1 | Model 2 | Model 3 | Model 4 |
| :---: | :---: | :---: | :---: | :---: |
| CLASS |  |  |  |  |
| Negative Climate | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Classroom Organization |
| Behavior Management | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Classroom Organization |
| Productivity | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Classroom Organization |
| Student Engagement | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Positive Climate | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Teacher Sensitivity | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Respect for Student Perspectives | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Instructional Learning Formats | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Content Understanding | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Analysis and Problem Solving | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Quality of Feedback | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| Instructional Dialogue | Ambitious Instruction | Ambitious General Instruction | Ambitious General Instruction | Ambitious General Instruction |
| MQI |  |  |  |  |
| Linking and Connections | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Explanations | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Multiple Methods | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Generalizations | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Mathematical Language | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Remediation | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Use Productions | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Student Explanations | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| SMQR | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| ETCA | Ambitious Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction | Ambitious Mathematics Instruction |
| Major Errors | Ambitious Instruction | Ambitious Mathematics Instruction | Mathematical Errors | Mathematical Errors |
| Language Imprecisions | Ambitious Instruction | Ambitious Mathematics Instruction | Mathematical Errors | Mathematical Errors |
| Lack of Clarity | Ambitious Instruction | Ambitious Mathematics Instruction | Mathematical Errors | Mathematical Errors |
| Number of Factors | 1 | 2 | 3 | 4 |
| Nested in | M2-M8 | M3-M8 | M4 |  |

Table 5b
Confirmatory Factor Analysis Model Organization for Bi-Factor Models

| Items | Model 5 | Model 6 | Model 7 | Model 8 |
| :---: | :---: | :---: | :---: | :---: |
| CLASS |  |  |  |  |
| Negative Climate | Ambitious Instruction | Ambitious Instruction | Classroom Organization | Classroom Organization |
| Behavior Management | Ambitious Instruction | Ambitious Instruction | Classroom Organization | Classroom Organization |
| Productivity | Ambitious Instruction | Ambitious Instruction | Classroom Organization | Classroom Organization |
| Student Engagement | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Positive Climate | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Teacher Sensitivity | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Respect for Student Perspectives | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Instructional Learning Formats | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Content Understanding | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Analysis and Problem Solving | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Quality of Feedback | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Instructional Dialogue | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| MQI |  |  |  |  |
| Linking and Connections | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Explanations | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Multiple Methods | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Generalizations | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Mathematical Language | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Remediation | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Use Productions | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Student Explanations | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| SMQR | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| ETCA | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction | Ambitious Instruction |
| Major Errors | Ambitious Instruction | Mathematical Errors | Ambitious Instruction | Mathematical Errors |
| Language Imprecisions | Ambitious Instruction | Mathematical Errors | Ambitious Instruction | Mathematical Errors |
| Lack of Clarity | Ambitious Instruction | Mathematical Errors | Ambitious Instruction | Mathematical Errors |
| Number of Factors | 3 | 4 | 4 | 5 |
| Nested in | M6-M8 | M8 | M8 |  |

[^3]Table 6
Model Fit Indices for Confirmatory Factor Analysis Models

| Fit Indices | Single Factor - No Cross Loadings |  |  |  | Bifactor - Items Load onto their Respective Instruments, Plus Other Factors |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 |
| Number of Factors | 1 | 2 | 3 | 4 | 3 | 4 | 4 | 5 |
| Number of Parameters | 75 | 76 | 78 | 81 | 101 | 102 | 102 | 104 |
| Akaike (AIC) | -1639.91 | -3128.08 | -3499.79 | -3795.80 | -3570.68 | -4000.36 | -3886.38 | -4059.54 |
| Bayesian (BIC) | -1342.45 | -2826.65 | -3190.43 | -3474.54 | -3170.10 | -3595.81 | -3481.83 | -3647.06 |
| Chi-Square Statistic | 3114.10 | 1912.41 | 1642.76 | 1399.24 | 1542.98 | 1199.75 | 1295.26 | 1169.947 |
| Chi-Square Degrees of Freedom | 275 | 274 | 272 | 269 | 249 | 248 | 248 | 246 |
| Scaling Correction | 1.223 | 1.212 | 1.183 | 1.172 | 1.183 | 1.162 | 1.164 | 1.1375 |
| RMSEA (Root Mean Square Error Of Approximation) | 0.163 | 0.124 | 0.114 | 0.104 | 0.115 | 0.099 | 0.104 | 0.098 |
| CFI | 0.484 | 0.702 | 0.751 | 0.795 | 0.765 | 0.827 | 0.810 | 0.832 |
| SRMR (Standardized Root Mean Square Residual) | 0.156 | 0.093 | 0.077 | 0.070 | 0.070 | 0.055 | 0.067 | 0.065 |
| Nested In | M2-M8 | M3-M8 | M4 |  | M6-M8 | M8 | M8 |  |

Draft Document. Please do not circulate or cite without permission.

Table 7a
Standardized Factor Loadings for CFA Model 4

| Items | Ambitious Mathematics Instruction | Ambitious <br> General Instruction | Classroom Organization | Mathematical Errors |
| :---: | :---: | :---: | :---: | :---: |
| CLASS |  |  |  |  |
| Negative Climate |  |  | 0.699*** |  |
| Behavior Management |  |  | $-0.841 * * *$ |  |
| Productivity |  |  | $-0.883 * * *$ |  |
| Student Engagement | $0.671^{* * *}$ |  |  |  |
| Positive Climate | 0.797*** |  |  |  |
| Teacher Sensitivity | $0.823 * * *$ |  |  |  |
| Respect for Student Perspectives | $0.821^{* * *}$ |  |  |  |
| Instructional Learning Formats | 0.673*** |  |  |  |
| Content Understanding | $0.831^{* * *}$ |  |  |  |
| Analysis and Problem Solving | $0.780^{* * *}$ |  |  |  |
| Quality of Feedback | $0.856^{* * *}$ |  |  |  |
| Instructional Dialogue | 0.886*** |  |  |  |
| MQI |  |  |  |  |
| Linking and Connections |  | 0.524*** |  |  |
| Explanations |  | $0.759 * * *$ |  |  |
| Multiple Methods |  | $0.523 * * *$ |  |  |
| Generalizations |  | 0.389*** |  |  |
| Mathematical Language |  | 0.368*** |  |  |
| Remediation |  | $0.575^{* * *}$ |  |  |
| Use Productions |  | 0.909*** |  |  |
| Student Explanations |  | 0.836*** |  |  |
| SMQR |  | 0.746*** |  |  |
| ETCA |  | $0.848^{* * *}$ |  |  |
| Major Errors |  |  |  | 0.834*** |
| Language Imprecisions |  |  |  | 0.508*** |
| Lack of Clarity |  |  |  | 0.876*** |
| Notes: $\sim \mathrm{p}<0.10,{ }^{*} \mathrm{p}<0.05,{ }^{* *} \mathrm{p}<0.01,{ }^{* * *} \mathrm{p}<0.001$ |  |  |  |  |

Draft Document. Please do not circulate or cite without permission.

Table 7b
Standardized Factor Loadings for CFA Model 8

| Items | Instrument Factors |  | Substantive Factors |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | CLASS | MQI | Ambitious Instruction | Classroom Organization | Mathematical Errors |
| CLASS |  |  |  |  |  |
| Negative Climate | -0.493*** |  |  | -0.486*** |  |
| Behavior Management | 0.451 *** |  |  | 0.836*** |  |
| Productivity | 0.619*** |  |  | 0.559*** |  |
| Student Engagement | 0.619*** |  | 0.237** |  |  |
| Positive Climate | 0.808*** |  | $0.100 \sim$ |  |  |
| Teacher Sensitivity | $0.795 * * *$ |  | 0.201** |  |  |
| Respect for Student Perspectives | 0.756*** |  | $0.333 * * *$ |  |  |
| Instructional Learning Formats | 0.615*** |  | 0.266*** |  |  |
| Content Understanding | 0.855*** |  | 0.078 |  |  |
| Analysis and Problem Solving | 0.719*** |  | 0.326*** |  |  |
| Quality of Feedback | 0.849*** |  | 0.180** |  |  |
| Instructional Dialogue | 0.820*** |  | $0.348 * * *$ |  |  |
| MQI |  |  |  |  |  |
| Linking and Connections |  | 0.573*** | 0.137 |  |  |
| Explanations |  | 0.903*** | 0.133 |  |  |
| Multiple Methods |  | 0.486*** | 0.245~ |  |  |
| Generalizations |  | 0.428*** | 0.084 |  |  |
| Mathematical Language |  | 0.382*** | 0.113 |  |  |
| Remediation |  | 0.704*** | 0.050 |  |  |
| Use Productions |  | 0.604** | 0.722*** |  |  |
| Student Explanations |  | 0.617*** | 0.578** |  |  |
| SMQR |  | 0.432* | 0.654*** |  |  |
| ETCA |  | 0.636** | 0.535* |  |  |
| Major Errors |  | $-0.238^{* * *}$ |  |  | -0.788*** |
| Language Imprecisions |  | $-0.166^{* *}$ |  |  | $-0.477 * * *$ |
| Lack of Clarity |  | -0.197** |  |  | -0.870*** |

Notes: $\sim \mathrm{p}<0.10,{ }^{*} \mathrm{p}<0.05,{ }^{* *} \mathrm{p}<0.01,{ }^{* * *} \mathrm{p}<0.001$


[^0]:    ${ }^{1}$ We note two important differences between instruments at the segment level. First, while the MQI has two raters score instruction, the CLASS only has one. Therefore, for the MQI, we averaged scores across raters within a given segment to match the structure of the CLASS. Second, while the MQI has raters provide scores for each seven-and-a-half minute segment, the CLASS instrument has raters do so every fifteen minutes. Therefore, to match scores at the segment level, we assigned CLASS scores for each fifteen-minute segment to the two corresponding seven-and-a-half-minute segments for the MQI.
    ${ }^{2}$ Multi-level bi-factor models did not converge.

[^1]:    ${ }^{3}$ To ensure that the resulting factor solutions were not affected by the differences in the scales used across the two instruments (MQI uses a three-point scale, whereas CLASS employs a seven-point scale), we ran the analyses twice, first with the original instrument scales and a second time collapsing the CLASS scores into a three-point scale (1-2: low, 3-5: mid, 6-7: high) that aligns with the developers' use of the instrument (see Pianta \& Hamre, 2009). Because there were no notable differences in the factor-solutions obtained from these analyses, in what follows we report on the results of the first round of analyses, in which we used the original scales for each instrument.

[^2]:    Notes: Descriptions of CLASS items from Pianta, Hamre, \& Mintz (2010).

[^3]:    Note: All models also include two method factors with all items cross loading onto their respective instrument factors.

